```
## Error in readChar(con, 5L, useBytes = TRUE): cannot open the
connection
```

```
## Error in eval(ei, envir):  object 'all_env' not found
```

Problem statement

We all want to do accurate Bayesian inference quickly:

- In terms of compute (wall time, model evaluations, parallelism)
- In terms of analyst effort (tuning, algorithmic complexity)

Markov Chain Monte Carlo (MCMC) can be straightforward and accurate but slow.

Black Box Variational Inference (BBVI) can be faster alternative to MCMC. **But...**

- BBVI is cast as an optimization problem with an intractable objective $\Rightarrow$
- Most BBVI methods use **stochastic gradient** (SG) optimization $\Rightarrow$
  - SG algorithms can be hard to tune
  - Assessing convergence and stochastic error can be difficult
  - SG optimization can perform worse than second-order methods on tractable objectives
- Many BBVI methods employ a **mean-field (MF) approximation** $\Rightarrow$
  - Posterior variances are poorly estimated

Our proposal: replace the intractable BBVI objective with a fixed approximation.

- Better optimization methods can be used (e.g. true second-order methods)
- Convergence and approximation error can be assessed directly

Outline

- BBVI Background and our proposal
  - Automatic differentiation variational inference (ADVI) (a BBVI method)
  - Our approximation: "Deterministic ADVI" (DADVI)
  - Linear response (LR) covariances
  - Estimating approximation error
- Experimental results: DADVI vs ADVI
  - DADVI converges faster than ADVI, and requires no tuning
  - DADVI's posterior mean estimates' accuracy are comparable to ADVI
  - DADVI+LR provides more accurate posterior variance estimates than ADVI
  - DADVI provides accurate estimates of its own approximation error
  - ADVI often results in better objective function values (eventually)
- Why don't we do DADVI all the time?
  - DADVI fails for expressive BBVI approximations (e.g. full-rank ADVI)
  - Pessimistic dimension dependence results from optimization theory
  - ...which may not apply in certain BBVI settings.

Notation

Parameter: $\theta \in \mathbb{R}^{D_\theta}$

Data: y

Prior: $\mathcal{P}(\theta)$ (density w.r.t. Lebesgue $\mathbb{R}^{D_\theta}$, nonzero everywhere)

Likelihood: $\mathcal{P}(y|\theta)$ (nonzero for all θ)

We will be interested in means and covariances of the (intractable) posterior

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(\theta, y)}{\int \mathcal{P}(\theta', y) d\theta'}.$$

Denote gradients with ∇, e.g.,

$$\nabla_\theta \log \mathcal{P}(\theta, y) := \left. \frac{\partial \log \mathcal{P}(\theta, y)}{\partial \theta} \right|_\theta \quad \text{and} \quad \nabla^2_\theta \log \mathcal{P}(\theta, y) := \left. \frac{\partial^2 \log \mathcal{P}(\theta, y)}{\partial \theta \partial \theta^\intercal} \right|_\theta$$

Assume we have a twice auto-differentiable software implementation of

$$\theta \mapsto \log \mathcal{P}(\theta, y) = \log \mathcal{P}(y|\theta) + \log \mathcal{P}(\theta).$$

Notation

Automatic differentiation variational inference (ADVI) is a particular BBVI method.

ADVI specifies a family $\Omega_{\mathcal{Q}}$ of D_θ-dimensional Gaussian distributions.

The family $\Omega_{\mathcal{Q}}$ is parameterized by $\eta \in \mathbb{R}^{D_\eta}$, encoding the means and covariances.

The covariances of the family $\Omega_{\mathcal{Q}}$ can either be

- Diagonal: "Mean-field" (MF) approximation, $D_\eta = 2D_\theta$
- Any PD matrix: "Full-rank" (FR) approximation, $D_\eta = D_\theta + D_\theta(D_\theta - 1)/2$

$$\underset{\mathcal{Q}\in\Omega_{\mathcal{Q}}}{\operatorname{argmin}} \mathrm{KL}\left(\mathcal{Q}(\theta|\eta)||\mathcal{P}(\theta|y)\right) = \underset{\eta\in\mathbb{R}^{D_\eta}}{\operatorname{argmin}} \mathrm{KL}_{\mathrm{VI}}(\eta)$$

$$\begin{aligned}\text{where } \mathrm{KL}_{\mathrm{VI}}(\eta) &:= \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\log \mathcal{Q}(\theta|\eta)\right] - \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\log \mathcal{P}(\theta, y)\right] \\ &= \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}}\left[\log \mathcal{Q}(\theta(z,\eta)|\eta)\right] - \underbrace{\underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}}\left[\log \mathcal{P}(\theta(z,\eta), y)\right]}_{\text{Typically intractable}}.\end{aligned}$$

The final line uses the "reparameterization trick" with standard Gaussian $z \sim \mathcal{N}_{\mathrm{std}}(z)$.

ADVI is an instance of the general problem of finding

$$\underset{\eta}{\operatorname{argmin}} F(\eta) \text{ where } F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}}\left[f(\eta, z)\right].$$

Two approaches

$$\text{Consider} \quad \underset{\eta}{\operatorname{argmin}} F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right].$$

Let $\mathcal{Z}_N = \{z_1, \ldots, z_N\} \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$, and let $\hat{F}(\eta|\mathcal{Z}_N) := \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n)$.

Algorithm 1
Stochastic gradient (SG)
ADVI (and most BBVI)

- Fix N (typically $N = 1$)
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - Draw $\mathcal{Z}_N$
  - $\Delta_S \leftarrow \nabla_\eta \hat{F}(\eta_{t-1}|\mathcal{Z}_N)$
  - $\alpha_t \leftarrow \mathrm{SetStepSize}(\mathrm{Past\ state})$
  - $\eta_t \leftarrow \eta_{t-1} - \alpha_t \Delta_S$
  - $\mathrm{AssessConvergence}(\mathrm{Past\ state})$
- **end while**
- **return** η_t or $\frac{1}{M} \sum_{t'=t-M}^{t} \eta_{t'}$

Algorithm 2
Sample average approximation (SAA)
Deterministic ADVI (DADVI) (proposal)

- Fix N (our experiments use $N = 30$)
- Draw $\mathcal{Z}_N$
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - $\Delta_D \leftarrow \mathrm{GetStep}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_{t-1})$
  - $\eta_t \leftarrow \eta_{t-1} + \Delta_D$
  - $\mathrm{AssessConvergence}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_t)$
- **end while**
- **return** η_t

Our proposal: Apply algorithm 2 with the ADVI objective.
Take better steps, easily assess convergence, with less tuning.

Experiments

For each of a range of models (next slide), we compared:

- **NUTS:** The "no-U-turn" MCMC sampler as implemented by PyMC [Salvatier et al., 2016]. We used this as the "ground truth" posterior.

- **DADVI:** We used $N =$ draws for DADVI for each model. We optimized using an off-the-shelf second-order Newton trust region method (`trust-ncg` in `scipy.optimize.minimize`) with no tuning or preconditioning.

Stochastic ADVI methods:

- Mean field ADVI: We used the PyMC implementation of ADVI, together with its default termination criterion (based on parameter differences).
- Full-rank ADVI: We used the PyMC implementation of full-rank ADVI, together with the default termination criterion for ADVI described above.
- RAABBVI: To run RAABBVI, we used the public package `viabel`, provided by Welandawe et al. [2022].

We terminated unconverged stochastic ADVI after 100,000 iterations.

Experiments

We evaluated DADVI on a range of models.

Model Name	Dim D_θ	NUTS runtime	Description
ARM (models)	Median (max)	median seconds (max minutes)	A range of linear models, GLMs, and GLMMs
Microcredit		minutes	Hierarchical model with heavy tails and zero inflation
Occupancy		minutes	Binary regression with highly crossed random effects
Tennis		minutes	Binary regression with highly crossed random effects
POTUS		minutes	Autoregressive time series with random effects

Table 1: Model summaries.

Experiments

```
## Error in eval(ei, envir):  object 'runtime_env' not found
```

Linear response covariances

Posterior variances are often badly estimated by mean-field (MF) approximations.

Take a variational approximation $\overset{*}{\eta} := \mathrm{argmin}_{\eta \in \mathbb{R}^{D_\eta}} \mathrm{KL}_{\mathrm{VI}}(\eta)$. Often,

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta] \approx \underset{\mathcal{P}(\theta|y)}{\mathbb{E}}[\theta] \quad \text{but} \quad \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Var}}(\theta) \neq \underset{\mathcal{P}(\theta|y)}{\mathrm{Var}}(\theta). \tag{1}$$

Example: Correlated Gaussian $\mathcal{P}(\theta|y)$ with ADVI.

Linear response covariances use the fact that, if $\mathcal{P}(\theta|y,t) \propto \mathcal{P}(\theta|y)\exp(t\theta)$, then

$$\left.\frac{d \underset{\mathcal{P}(\theta|y,t)}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \underset{\mathcal{P}(\theta|y)}{\mathrm{Cov}}(\theta). \tag{2}$$

Let $\overset{*}{\eta}(t)$ be the variational approximation to $\mathcal{P}(\theta|y,t)$, and take

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \left.\frac{d \underset{\mathcal{Q}(\theta|\overset{*}{\eta}(t))}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right) \left(\nabla^2_\eta \mathrm{KL}_{\mathrm{VI}}\left(\overset{*}{\eta}\right)\right)^{-1} \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right)$$

Example: For ADVI with a correlated Gaussian $\mathcal{P}(\theta|y)$, $\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Cov}}(\theta)$.

Linear response covariances

Posterior variances are often badly estimated by mean-field (MF) approximations.

Take a variational approximation $\overset{*}{\eta} := \mathrm{argmin}_{\eta \in \mathbb{R}^{D_\eta}} \mathrm{KL_{VI}}(\eta)$. Often,

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta] \approx \underset{\mathcal{P}(\theta|y)}{\mathbb{E}}[\theta] \quad \text{but} \quad \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Var}}(\theta) \neq \underset{\mathcal{P}(\theta|y)}{\mathrm{Var}}(\theta)\,. \tag{3}$$

Example: Correlated Gaussian $\mathcal{P}(\theta|y)$ with ADVI.

Linear response covariances use the fact that, if $\mathcal{P}(\theta|y,t) \propto \mathcal{P}(\theta|y)\exp(t\theta)$, then

$$\left.\frac{d\, \underset{\mathcal{P}(\theta|y,t)}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \underset{\mathcal{P}(\theta|y)}{\mathrm{Cov}}(\theta)\,. \tag{4}$$

Let $\overset{*}{\eta}(t)$ be the variational approximation to $\mathcal{P}(\theta|y,t)$, and take

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \left.\frac{d\, \underset{\mathcal{Q}(\theta|\overset{*}{\eta}(t))}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right) \left(\nabla^2_\eta \,\mathrm{KL_{VI}}\left(\overset{*}{\eta}\right)\right)^{-1} \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right)$$

Example: For ADVI with a correlated Gaussian $\mathcal{P}(\theta|y)$, $\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Cov}}(\theta)$.

J. Salvatier, T. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016. doi: 10.7717/peerj-cs.55. URL https://doi.org/10.7717/peerj-cs.55.

M. Welandawe, M. Andersen, A. Vehtari, and J. Huggins. Robust, automated, and accurate black-box variational inference. *arXiv preprint arXiv:2203.15945*, 2022.